C

Chaty

Private local AI chat, knowledge base, deep research, and voice for Mac and Windows.

About

Chaty is a local AI chat app that runs open LLMs entirely on your own machine. It is built for privacy first: no account, no cloud, and no telemetry, with conversations kept on disk instead of sent to a server. Chaty supports chat, a private knowledge base, multi-round Deep Research, and voice, all offline on Windows and macOS.

What is Chaty

Chaty is a local AI chat app for running open LLMs on your own computer. It is designed as a private AI workstation rather than a hosted chatbot, so the model, your prompts, and your files stay on-device. Chaty supports standard chat, a private knowledge base,

The app is built around local inference. You can load any `.gguf` model, and on Apple Silicon it also supports MLX models natively. That makes Chaty useful if you want an AI tool without accounts, cloud syncing, or server-side logging.

Key features

Chaty combines several local AI workflows in one desktop app:

- **Local model support:** load `.gguf` files and, on Apple Silicon, MLX model folders.
- **Private operation:** no account, no cloud dependency, and no telemetry.
- **GPU acceleration:** Vulkan on Windows and Metal on Apple Silicon can offload work to the GPU, with CPU fallback when needed.
- **Persistent context:** reused KV-cache and saved conversation state help keep multi-turn sessions fast.
- **Auto-fit context management:** the app can fit larger chats by summarizing non-destructively when context gets long.
- **Code mode / agent workflow:** point Chaty at a folder and let it inspect files, edit code, search, and run commands with live task steps.
- **Knowledge base / RAG:** documents such as PDFs, Markdown, code, and scanned images can be indexed locally.
- **Deep Research:** the app can plan queries, search in multiple rounds, and produce a cited report.
- **Voice mode:** you can talk to the app hands-free.

Chaty also exposes sampling controls and saveable prompt presets, which is helpful if you work with the same model settings often.

How to use it

1. Download Chaty for Mac or Windows from the official site.
2. Choose a local model in `.gguf` format, or use an MLX model on Apple Silicon.
3. Open a chat and start prompting; the app handles the model template and tokenizer from the model itself.
4. If you need grounded answers, add files to the knowledge base and limit the search

For longer investigations, switch to Deep Research, enter a topic, and review the search and writing stages as they happen.

If you are coding, toggle to Code mode, point the app at a workspace folder, and approve changes step by step.

Use voice mode when you want a spoken back-and-forth instead of typing.

Chaty is most useful when you already have local model files or know where to get them, because the app is built around running those models on your own hardware.

Who it is for

Chaty is a good fit for people who want a local AI chat app with strong privacy guarantees. It is especially relevant for:

developers who want a coding agent that stays inside a local workspace

researchers who need a knowledge base with cited, source-grounded answers

users who prefer offline tools and do not want an account

Mac users with Apple Silicon who want native MLX support

Windows users who want GPU-accelerated local LLMs without cloud services

It is less useful for someone who wants a simple hosted chatbot, because Chaty is centered on local models and local file access.

What to know before you use it

Chaty is powerful, but it is not a hosted model service and it does not magically provide model quality on its own. You are responsible for supplying compatible local models, and performance will depend on your hardware, VRAM, and model size. The app notes GPU offload with a CPU fallback, which means slower systems can still run it, but not necessarily at the same speed as a strong GPU setup.

The knowledge base is grounded in the files you provide, but it is still an AI retrieval system, so it can miss relevant passages if the source material is incomplete or poorly structured. Deep Research searches the web and cites what it actually used, but for critical facts, official documentation and primary sources are still the safer reference. The same applies to code assistance: Chaty can inspect and edit files locally, yet you

If you need a private AI chat app, a local AI chat app, or a desktop AI assistant for offline work, Chaty is a focused option with unusually broad local workflows. It is strongest when you want AI chat, knowledge base retrieval, Deep Research, and voice in one privacy-first local package.

FAQ

- Does Chaty really work offline?
- Is Chaty free to use?
- What models can I use with Chaty?
- Does Chaty work offline?
- Can I code with Chaty?
- What platforms does Chaty support?
- How does the knowledge base work?
- Can Chaty use my own documents?
- Is Chaty good for coding tasks?

Sponsored

Belay – Awa... Featured
Developer Tools

Partly Featured
Finance

View ↗

Free # Finance

View ↗

Visit website ↗

chaty

Pricing	Free
Category	AI
Platform	Mac, Windows
Launched	Aug 2026
Updated	15 days ago
Verified	1 months ago

TAGS

local # privacy # rag # llm # local-ai # voice # deep-research # offline # open source # code-agent

Contact

Terms

© 2026 Dir Hub

Curated indie apps

C Chaty